DOCUMENT RESUME

EA 001 969

ED 026 732

By-Welty, Gordon A.

The Logic of Evaluation.

Pub Date Oct 68
Note-29p.
EDRS Price MF-$0.25 HC-$1.55
Descriptors-*Decision Making, *Evaluation Criteria, *Feedback, Literature Reviews, *Methodology, *Program Evaluation

The logic of the evaluation of educational and other action programs is discussed from a methodological viewpoint. However, no attempt is made to develop methods of evaluating programs. In Part I, the structure of an educational program is viewed as a system with three components--inputs, transformation of inputs into outputs, and outputs. Part II discusses the necessary condition for a program to be a system (the presence of feedback loops) citing as one example the school system with an evaluation unit. In Part III, the possibility of mapping experimental designs into social space characterized by feedback loops is confirmed while refuting statements by Stufflebeam to the contrary. Part IV examines the historical precedence for the findings and concludes that it is possible, from a methodological viewpoint, to implement a rigorous experimental design and also to provide feedback for managerial decisionmaking in the context of action research. (HW)

ED026752

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

The Logic of Evaluation

by Gordon A. Welty
Chatham College

Educational Resources Institute
October 1968

EA 001 969

Acknowledgments

The work incorporated in this essay has occupied my attention for several years, and has benefited during that time from the criticism of many persons. First and foremost has been my mentor, Professor E. Grunberg of Akron, who first turned me towards the general problem of reflexive behavior on which he has signally worked, and indicated the means for solving some of the problems associated with reflexive behavior. Next, Dr. Donald Henderson of St. Louis showed me the specific problem related to experimental design and enabled me to work on aspects of it. Ronald Wilkes provided valuable inputs, as did M.J. Duda and Dr. D.K. Stewart, during my last and most intense year's work in Pittsburgh. Mrs. Duda has never ceased in the support she provided me. Mr. Alex Ajay has always reinforced my belief in the value of criticism, and also introduced me to the work of Russell Ackoff. Laurie Dancy, Judy McBroom, and Richard Fogel have provided much needed assistance in clarifying my thoughts. I must also thank Dr. M.P. Provus, who has provided me with time to work on some facets of this problem. Finally, Professor Alan Anderson of Pittsburgh might be mentioned as one who provided no assistance in these matters, when assistance was requested at a crucial formative stage of problem-solving. This latter may be a reflection on higher education in America.

Of course, none of these persons are responsible for errors remaining, except to the extent that I was allowed by them to disregard their counsel.

Introduction

The following essay discusses the logic of the evaluation of educational and other action programs. As such, it is a methodological statement, introducing methods only incidentally. If one seeks to discover methods of evaluating programs here, they will be disappointed. For that the reader must look elsewhere. Only formal problems of the possibility of evaluation are treated. In fact, this essay has so little regard for methods that we have not at all concerned ourselves with constructive arguments (in the sense that our proofs will not lead to algorithms).

This document need not be justified further: anyone who knows the literature on evaluation is aware of the terrible conceptual muddle of educational project assessment. Those who don't know the literature should have stopped reading a paragraph ago. The essay consists of four parts. Part I gives the structure of an educational program. Part II discusses the necessary condition for the program to be a system, the presence of feedback loops. Part III discusses the possibility of mapping experimental designs into social space characterized by feedback loops. Part IV examines the historical precedence for the findings reported.

I.

Theorists in the social sciences and education are thinking more and more of their subject matter in terms of "systems," to which, as to any system, process and adaptive control are essential. The following is an examination of such a system and a discussion of several of its relevant properties.

Three things are essential to any process: first, inputs or raw materials, second, a transformation to convert the raw material, and third, the output or finished product. In an educational system, for example, the student might be the input, the system the combination of curriculum, teachers, and physical plant, etc., and the output the high school graduate.

A means of guaranteeing that the system will produce the desired output must also be provided. If society wants more electrical engineers, for example, it does not want electronic technicians. Probably either job could be performed by the same individual. The job he does perform depends, in large part, upon the standards of the educational system of which he is the product. One set of standards or criteria, by specifying the level of abilities and competencies required for t job, defines an electrical engineer, while a second set of criteria defines the technician. When the individual can meet one of these sets of criteria, he can fill the specified job.

Criteria are established, then, in order to define and permit control

of the system. When it is noticed that the criteria are not being met, appropriate changes are made in the program to reestablish accord with the criteria and eliminate the discrepancies. This requires a feedback or control loop as a formal property of the system. It should be emphasized that there are two ways the violation of standards can be characterized: either the individual product is "defective," or the system is inadequate to its task.

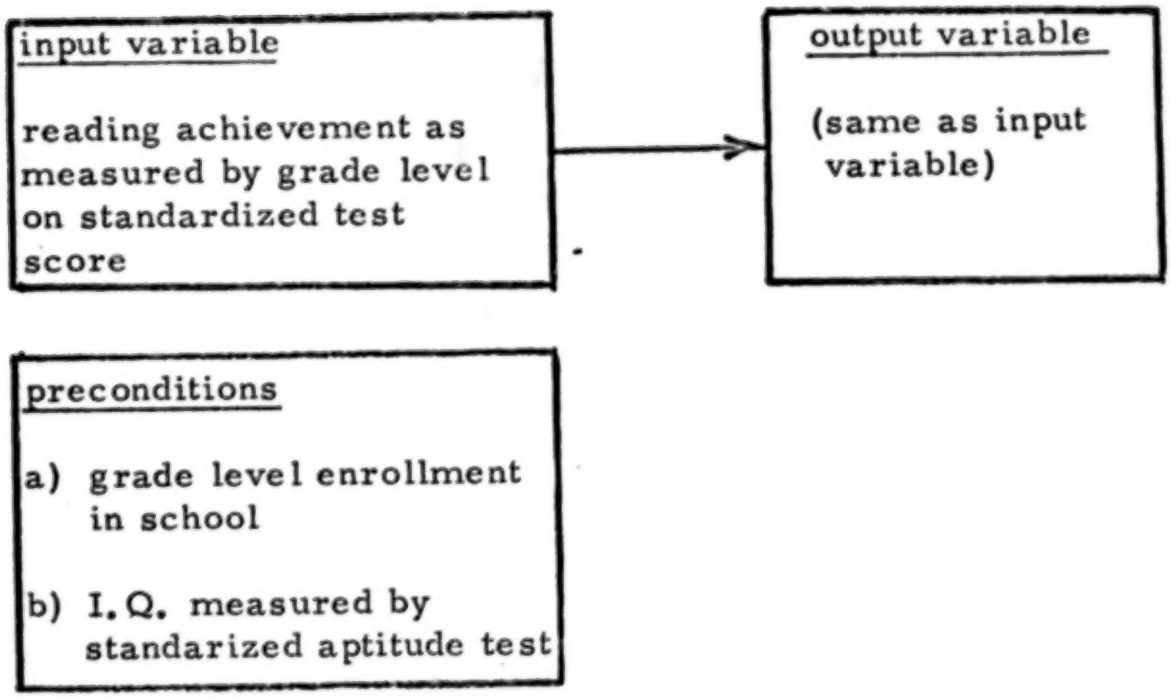
Within the system are a number of programs designed to achieve the specific goals of the system. It is possible to schematize a program completely by a consideration of the characteristic kinds of behavior involved in that program. These kinds, or dimensions, of behavior fall into three classes: "input variables," i.e., a set of dimensions of behavior which exists upon the subject's entry into the program, and which will be changed by the action of the program; "output variables," i.e., a set of dimensions of behavior, identical to the input variables, existing at the point of exit from the program as the result of the program's action on the input variables, and "preconditions," a set of dimensions of behavior which is associated with the input and output variables but which will be unaffected by the program. In fact, the collection of dimensions of behavior indicated here defines a vector space. This accounts for the equal number of input and output dimensions (by the Principle of Dimensional Homogeneity). This becomes important at a later stage of our discussion, when we consider program change.

A compensatory program in the Pittsburgh Public Schools titled the "Transition Room Program" affords a good example.* The purpose of the Transition Room is to help underachieving children solve their reading problems before they enter the fourth or fifth grade. Up to the fourth and fifth grades, learning to read is an end in itself. In these grades, however, it becomes a means to the acquisition of knowledge in other substantive areas: a "transition" is made from reading as subject matter to reading as a communication skill. The Transition Room Program is designed to facilitate this transition. In order to reach those children most in need of aid, selection criteria have been set up: children entering the program must have MAT reading achievement scores at least one year below grade level, indicating underachievement; must have an I.Q. of 85 or above, indicating a capacity for benefiting from remedial instruction; and must be enrolled in third or fourth grade. The goal of the program is to raise the MAT score to grade level.

When considered in terms of dimensions of behavior, the program may be broadly described by the following diagram:

---

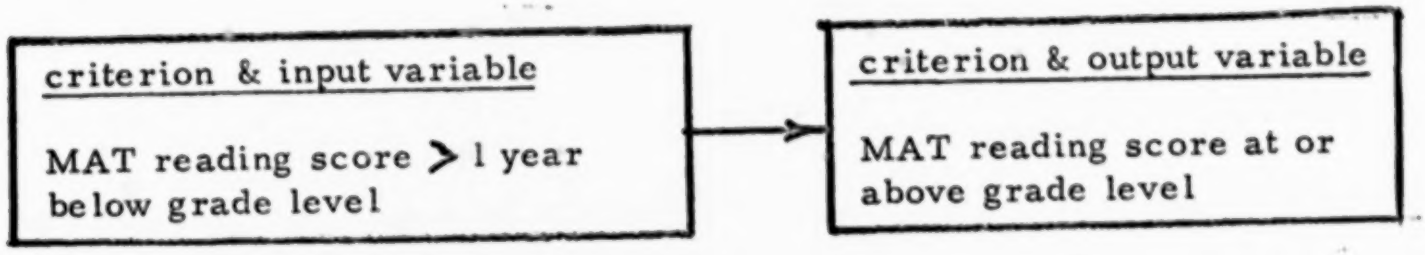
* This example is discussed further in my "Evaluation of Public School Programs," a paper presented at L.R.D.C. in Pittsburgh 18 November 1968.


input variable
reading achievement as measured by grade level on standardized test score
output variable
(same as input variable)
preconditions
a) grade level enrollment in school
b) I.Q. measured by standarized aptitude test

In the case of the Transition Room, there is only one input and associated output variable. In other programs there may be several, each to be acted upon by the program to produce an associated output variable. For each pair of change variables (that is, for each input-output pair) there is one process to transform the value of the input dimension to the value of the output dimension. The description of the process involved in the program may be made more specific if it is borne in mind that it is necessary to find a condition sufficient to effect the change from input to output for each pair. This becomes clear with consideration of criteria.

Criteria come into existence when we specify thresholds or ranges of values for each dimension of behavior. Specifying values for the input and precondition variables provides a description of selection criteria.

Specifying values of the output variable provides a description of the goals of the program. To take the Transition Room example again,


criterion & input variable
MAT reading score > 1 year below grade level
criterion & output variable
MAT reading score at or above grade level
criterion & precondition
a) 3rd or 4th grade enrollment
b) IQ ≥ 85 (Stanford-Binet L-M)

Further consideration of this example raises another point of interest. At the end of two school years, the student leaves the Transition Room because he no longer fulfills the precondition of third or fourth grade enrollment, whether his MAT score is at grade level or not. This points out the existence of an output variable that has no criterion associated with it; that is, there are states of the program, at termination, independent of goal achievement. Before goal discrepancies are evaluated a way must be found to characterize this terminal state, and as we can see, if the output state is described independently of goal state, but described in terms of the same dimensions, it becomes possible to characterize these discrepancies.

II.

The type of program that corresponds to the schema presented above is the "open-loop" control or implicit system. The next approximation

to the complexities of the educational system is the simple closed-loop control system, or "feedback" control system: for example, the school system with an evaluation unit. This is what can be called an "action program," and is explicitly a system.

Given a continuous evaluation activity, system control becomes essentially a statistical problem. In quality control of, say, ball-bearings, the steps would be, first, selection of variables (our input-output "pair") describing the materials,second, specification of parameters to provide criteria defining both the acceptable product and the acceptable functioning of the process, and then comparison of the product and process with the criteria. For example, if the production line transforms Babbitt metal and steel into ball-bearings, where diameter and weight are the descriptive variables, measures of central tendency and dispersion (the mean and standard deviation) will be specified in order to determine the tolerable amount of dispersion. The mean constitutes the expected value: if the weight of the ball-bearing is to be 3 grams, then ideally the weight of each and every ball-bearing will be 3 grams (of course measured on a perfect balance). Incidentally, preconditions might well be included: the specific gravity, tensile strength, etc. of the materials might be indicated to fall in a specified range.

Of course, the quality control engineer never attempts to make the observed distribution, derived from the measurement of the variables selected as descriptive of the material, <u>exactly</u> coincide in the parameters, with the expected distribution. Any induced change in the observed

distribution, or output measure, that is, any action to make the product conform more closely with the criteria, must be considered a cost and subtracted from the cost incurred by the number of rejects in the process. Such action must be undertaken only to minimize overall cost (while maintaining output constant).

The two key elements in quality control appear, then, to be the specification of the expected values of the variables selected as descriptive of the product, and the measurement of the actual material or substance undergoing the process in order to ascertain whether the product is exceeding the tolerable deviation from the specified values. We have seen that the decision to take action as a result of these actual measurements rests on considerations of efficiency.

Consequently, rather than the simple linear equation

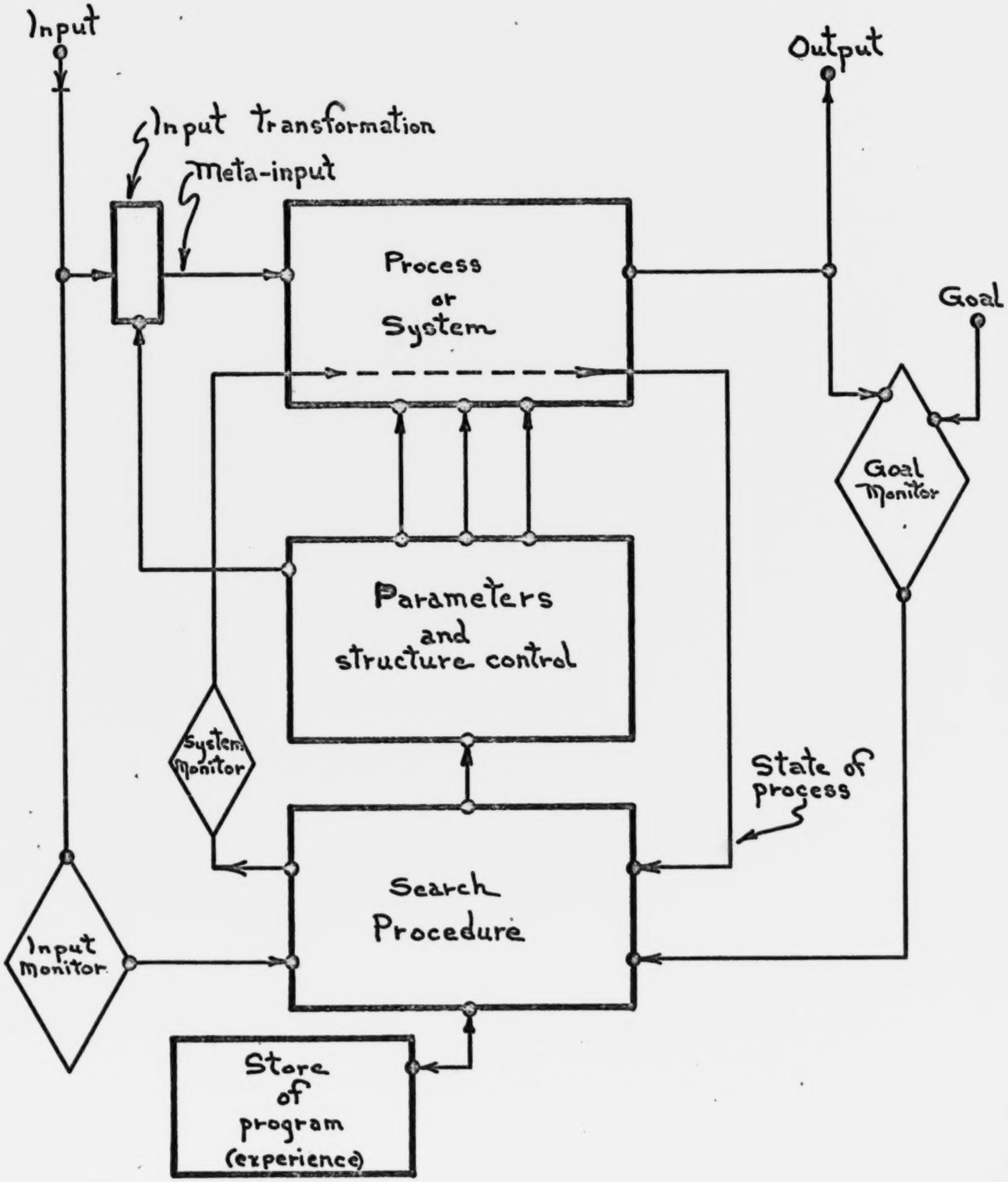
$$y = ax + b$$

which describes the open-loop system, we have two independent variables, x and the parameter a, where a is a function of the goals. Thus we have

$$y = f(g)x + b$$

where f(g) is the statement of a as a function of goals g (See System "S" following). But how does one find the values defined by the criteria?

A viable technique that we have used in Pittsburgh can be called the "value-finding" approach. As an example of this, a meeting was held of the staff members of a program. At that time they were asked to rank order six possible objectives of the program. This data was analyzed on



SYSTEM "S"

Adopted from B. McR. Sayers "Self-Organizing Systems," in J.H. Westcott, ed. An Exposition of Adaptive Control (Pergamon, 1962).

the spot and the results were presented to the staff members later in the meeting. They were than asked to (1) comment on the points of consensus and (2) explain why they felt the points of dissensus existed. Consensus could be considered a manifestation of a value.

The analysis proceeded as follows: The participants ranked the objectives from one to six. These "votes" were then arrayed in a 6 x 6 table, with the objectives to be ranked comprising the rows, and rank-orders one to six comprising the columns. The frequency with which a given objective was given a specified rank was the cell entry. Variance could then be estimated. A semi-interquartile range of one rank or less was said to indicate consensus, and a range of more than one, dissensus.

RATINGS

		1	2	3	4	5	6
Items	1	4	5	2	1		
	2	5	2	2	2	1	
	3	1	3	5	3		
	4	2	2	1	5	2	
	5			1		8	3
	6			1	1	1	9

The modal values have been underlined. This gives us a good notion of central tendency. A rough and ready notion of variance is provided by the semi-interquartile range (SIQ R: 25th to 75th centile). For the six items, it was as follows:

Items	SI QR
1	1+
2	2-3
3	1+
4	2
5	1-
6	1-

The smaller SIQR means the lesser variance. Maximal interrater agreement (i.e. a "value") would be indicated if each item has a SIQR of less than 1. Notice that both items 2 and 4 had large variances. In the case of 2, it was decided that the item (a) was ambiguous, and (b) was a supervenient, not terminal goal of the program. Item 4 was probably pretty vague too. A high variance is generally indicative of vagueness (lack of reliability).

Of interest is the unanimity concerning items 5 and 6, which were disvalues. There is an obvious difference between items 1-4, and 5-6.

When the results were communicated to staff at the same meeting, it was suggested that the values expressed conformed to the structure of the program. The program has been established with two somewhat incompatible major goals. The various objectives related to one of these major goals were values, rated high. The items related to the other major good were values, rated low. There had been included in the list of objectives two supervenient objectives, which might have provided the basis of resolution of the incompatabilities, between the major goals. Those items which related to the supervenient objectives, showed no consensus. They were found on inspection by the staff to be vague and amorphous.

III.

Speaking of evaluation in general, C.I. Lewis points out that actions could not attain success except that there are evaluations, which are essentially predictions.[1] "Whether the action is performed or not will depend upon evaluations made."[2] In terms of general systems theory, Lewis is emphasising the necessity of feedback in any action scheme.

Action is an attempt to control the future, as far as is possible, for our own benefit. Action is based in the present, in the given situation; it is intentional behavior directed towards realizing a desirable state-of-affairs, or avoiding an undesirable state-of-affairs.[3] The movement is from the reality of the present to a chosen future. Lewis continues that "the principal function of empirical knowledge is that of an instrument enabling transition from the one to the other."[4]

---

[1] C.I. Lewis, An Analysis of Knowledge and Valuation (1946), pp. 371-372.

[2] ibid. p. 4

[3] ibid. ch. XII, esp. p. 367 f.

[4] ibid. p. 4

Feedback of evaluative reports to decision makers is a necessity in the rationally managed school system, if it is viewed as a system. On the revision of ongoing educational programs, for instance, Hastings has stated that

> without such feedback, either the decision to revise or the decision not to revise--and most certainly the decision of how to revise--must be based upon feeling tones and the arguments of personal preference.[5]

It has been frequently maintained that the demand for evaluative feedback is incompatible with the classical experimental design. For example, consider Stufflebeam's statement that "the experimental design type of evaluation prevents rather than promotes changes in the treatment."[6] He suggests that a necessary condition for the implementation of the classical design is

> treatment and control conditions must be applied and held constant throughout the period of the experiment, i.e. they must conform to the initial definitions of these conditions.[7]

The conditions of invariance of treatment and control are sufficient then, on Stufflebeam's argument, to preclude program change. Thus, these

---

[5] J.T. Hastings, "Curriculum Evaluation - The Why of Outcomes," Journal of Educational Measurement, III: 1(1966), p. 27.

[6] D.L. Stufflebeam, "Evaluation as Enlightment for Decision-Making," Evaluation Center, OSU College of Education (mimeo, 1968), p. 13

[7] Stufflebeam, op.cit., p. 12

conditions are sufficient to preclude evaluation feedback for managerial decision-making, because "the new or traditional program conditions could not be modified in process, since in that event one could not tell what was being evaluated." [8]

Stufflebeam considers this a "problem relating to the methodology of evaluation." [9] As such, it is a problem of paramount importance to the development of evaluation theory. If the dissemination of evaluative findings is not permitted, to preserve the classical design, then evaluation loses its value to the decision-maker. On the other hand, if the classical design is to be abandoned, serious problems await evaluators in the development of an alternate.

We find, however, that these contentions are not valid. It is not the case that treatment must remain invariant. Corrective action by program managers, in light of evaluative feedback, can take place concurrent with an evaluation in the framework of the classical experimental design.

Of crucial importance is what Stufflebeam intends by "changes in the treatment." To suggest that random treatment variance is included here is hardly acceptable. [10] The question might be raised: if quantitative change is a change of the value of a given variable, (whether an intensive or extensive measurement) then qualitative change is a change of a variable

---

[8] Stufflebeam, ibid.

[9] op. cit., p. 11

[10] Among other things, this suggestion violates Wold's Theorem of predictive decomposition.

or dimension itself. Could not the latter be Stufflebeam's intent? Of course, if by change of a variable or dimension is meant movement in space of a function by an operator, then there is no difference between quantitative and qualitative change. This is clear, given that the operator is known. However, let us consider the other possible meaning of qualitative change: a variable or dimension is simply added or deleted from the analysis, and to this "changes in treatment" corresponds.

The Principle of Dimensional Homogeneity states that for a given equation, all the dimensions or variables in the equation can be categorized in terms of a collection of fundamental measures. For example, if volume occurs in an equation, the dimensions of volume are categorized in terms of length. The principle also states that the dimensionality of the variables (the dimensionality of volume is 3) on the right- and left-hand side of the equation, by fundamental measure, must be equal.[11] For example, velocity is distance per time. That is, $v = d/t$. The fundamental measures for velocity are two, length and time. Dimensionality is 1 and -1, respectively. As this also is the case for distance divided by time, the formula is dimensionally homogeneous. If it were otherwise, the introduction or deletion of either fundamental measures or dimensionality across the equation would be a case of ad hoc theorizing (however subtle).

---

[11] This is known as the π Theorem.

So it is not possible, as a methodological (not necessarily ontological) point, for there to be qualitative change in this sense. Hence we consider here only the case of treatment variance which is both a rational response to evaluative feedback and also dimensionally homogeneous, and indicate how this is compatible with the classical design.

We can represent a project ϕ, to be evaluated, as

$$x_o = \phi\,(x_i);$$

in an n-dimensional vector space x, the vector x_i is a measure of inputs, x_o, a measure of outputs, and ϕ is the transformation or process. As we are speaking of an action program, this is sufficient on Lewis' terms for continuous monitoring and feedback of the project's state. Note that this continuous feedback is a part of the project.

Let a research design be given by the structural equation

$$x' = \overline{x} + p + \epsilon$$

for x', the criterion measure for the experimental group; $\overline{x}$, the population mean; p, the effect of treatment; and ϵ the experimental error, a randomly distributed independent variate of zero mean. As evaluative feedback on treatment is continuous throughout the evaluation cycle, we can restate this as the dynamic equation

$$x' = P\,(\overline{x}) + \epsilon$$

where P is the specific treatment incorporated in process ϕ.

Now if $x' \neq \overline{x}$ is significant there is a treatment effect. Further, dissemination of evaluative findings as expectancies will be labelled E(x). Thus for $E(x') = E(\overline{x})$, the rational manager will react by changing

P to P*, effecting program improvement, and, on Stufflebeam's argument, rendering invalid the evaluation within the classical design.

However, let us introduce a reaction function R describing the rational manager's response:

$$P^* = R(P)$$

The contractive mapping theorem tells us, given a metric space $\langle P, d \rangle$, for any reactions RP_j, RP_k, if

$$d(RP_j, RP_k) \leq \alpha d(P_j, P_k)$$

where α is a contractive constant (the economists "coefficient of expectations" or "adjustment") then there exists a unique fixed point π such that $\pi^* = \pi$. Here the reaction functions becomes an identity transformation, i.e. the manager has ceased to change the program in response to the feedback. The value π is the state of process stability, hence the correct evaluative judgment x_o^* will be

$$x_o^* = \pi(x_i)$$

Given adequate resources, this will guarantee that the desired significant difference $x_o^* \neq \overline{x}$, is realized.[12] We find here sufficient conditions for the falsity of Stufflebeam's methodological argument.

---

[12] This argument is available in my "Experimental Designs and Applied Research" California Journal of Educational Research" (1969) and "The Use of Experimental Designs in the Decision-Making Feedback Process," Journal of Experimental Education (1969).

IV.

In the writings of Fisher, we find sufficient conditions for the methodological (again not practical) arguments for the use of experimental designs in evaluative research.

In his Design of Experiments, Fisher proposes to "examine the physical conditions of the experimental technique."[13] After mentioning that matching of conditions across treatment levels in the experimental design is a formal condition for minimizing errors, Fisher argues it is impossible to realize this condition in fact, since "uncontrolled causes which may influence the result are always strictly innumerable."[14] With regards matching of conditions, the assumption that "refinements constitute improvements to the experiment" is dismissed on the basis of cost considerations. Since matching is a sufficient but not a necessary condition, control of errors in the experiment can and must be realized by other means. The cost of complete matching across treatment levels would be (quite strictly) infinite, and since "an essential characteristic of experimentation is that it is carried out with limited resources," Fisher proposes randomization as an alternative. This is a procedure by which the experiment "may be guaranteed against corruption by the causes of

---

13 Sir R.A. Fisher, Design of Experiments, 8th ed. (1966), p. 17

14 Fisher op.cit., pp. 17-18

disturbance which have not been eliminated." [15] Thus there are two and only two sufficient conditions for experimental control; hence, one of the two is always necessary. Irrelevant variables are eliminated in effect either by matching of conditions ("eliminated in the field") or by randomization.[16] Fisher emphasizes the sufficiency of the latter technique when he argues that

> it is apparent that the random choice of the objects to be treated in different ways would be a complete guarantee of the validity of the test of significance, if these treatments were the last in time of the stages in the physical history of the objects which might affect their experimental reaction.[17]

This is to say that randomization is sufficient in the absence of treatment variation, to which Stufflebeam would undoubtedly agree.

In evaluation or action research, a new aspect is added. Because of the various institutional contingencies, it is usually an unacceptable policy to randomly choose subjects for treatment. It would be possible, for instance, to take the lower two-fifths of the students, as ranked by a standardized achievement test. This group could then have remedial treatment provided, by random assignment, to one half, which would

---

15 op. cit., p. 19. More precisely, random assignment of subjects to treatment levels permits a precise estimate of error.

16 Sir. R.A. Fisher, "The Arrangement of Field Experiments," Journal of Ministry of Agriculture (1926), p. 509.

17 Design of Experiments, p. 20.

amount to one-fifth of the total population. However, it is usually policy to take the lowest fifth, and administer treatment to them as a group. Thus no "control group" is available. This is, however, an institutional contingency, hence not a methodological problem per se. Randomization is still a possibility, hence Fisher's discussion of randomization is relevant to the methodology of evaluation.

Fisher generalizes his argument at this point by pointing out that variance in treatment subsequent to randomization presents no "practical inconvenience." He states

> subsequent causes of differentiation, if under the experimenter's control...can either be predetermined before the treatments have been randomized, or, if this has not been done, can be randomized on their own account. [18]

The first alternative here is merely the recognition that the rational decision maker's response to evaluative feedback is program change. The second alternative is excluded from our discussion, as randomly distributed response by a program manager is not conducive to systematic pursuit of policy.

At this point, we can discuss three possible sources of error: (a) consequences of differences already randomized which are accounted for by the initial randomization, (b) natural consequences of the difference in treatment levels; since the null hypothesis argues there will be no

---

[18] ibid.

treatment effect, there can be no consequences of this effect, and (c) effects supervening by chance, independent of treatment levels; because of random assignment, estimates of deviance from a specified distribution across all treatment levels for these effects can be given. Any systematic variance will have been eliminated by the initial randomization.

The dissemination of evaluation findings to the rational program manager will produce program change. As a corollary of the Principle of Management by Exception, we know that if a defect in the program is noted and reported, given adequate program resources, the defect will be corrected by the rational manager. Thus both the corrective action of the manager, and the adequacy of resources are determinate. As such, the variance of treatment as a function of evaluative feedback can be, in Fisher's terms, "predetermined." The crucial issue is whether we try to explain the world exclusively, as Toulmin and Goodfield put it, on the analogue of the "sixteenth century, or even medieval, machines," or whether we view things in terms of "twentieth century machines."[19] And this is a sufficient condition for the compatibility of the classical experimental design and the dissemination of evaluative findings.

---

[19] S. Toulmin and J. Goodfield, The Architecture of Matter (1962), p. 334. See also K. Boulding "General Systems Theory - The Skeleton of Science" in Educational Data Processing ed. R.A. Kaimann and R. W. Marker (1967) pp. 6-15, levels ii, iii and iv; also H. Goode, Ch. 6 in Systems: Research and Design ed. D. P. Eckman (1961) pp. 105-117.

Hence we see that, contrary to the contentions of Stufflebeam it is possible to implement a rigorous experimental design, and also provide feedback for managerial decision-making, in the context of action research. Whether practical concerns, such as the competence of the researcher, or the resources and administrative support available to him, do in fact militate against his ability to implement a rigorous design, is not a methodological issue, and not under consideration here. On the other hand, if the manager is oblivious to feedback, or responds to feedback with random and affective behavior rather than systematic and rational action, this is a psychological issue, and not under consideration here. But the methodological "problem," posed by Stufflebeam, can be considered ill-conceived and non-existent.

APPENDIX

We have included here a representative collection of quotations. The authors cited here are addressing themselves to the conjectured incompatibility of experimental designs and evaluation or action research. They are quoted in alphabetical order.

Brooks, Michael P. "The Community Action Programs as a Setting for Applied Research," Journal of Social Issues (1965), p. 38.

> ...continuous feedback of research findings into community action programs, thereby producing adjustments and improvements in their operation... has the unfortunate effect of tossing a monkey wrench into the research design constructed at the program's outset.

Dyer, Henry S. "Overview of the Evaluation Process" On Evaluating Title I Programs. Educational Testing Service, Princeton, New Jersey (1966), p. 18

> We evaluate, as best we can, each step of the program as we go along so that we can make needed changes if things are not turning out well. This view of evaluation may make some of the experimental design people uneasy because it seems to interfere with the textbook rules for running a controlled experiment... There is one kind of evaluation to be used when you are developing an educational procedure...I would call concurrent evaluation. And there is a second kind of evaluation...I would call ex post facto evaluation; it is what the experimental design people are usually talking about when they use the word evaluation.

Guba, Egon G. "Confronting the Problems of Educational Evaluation: A Call for a Consortium of Relevant Agencies," NISEC, Bloomington, Indiana (mimeo, Dec. 1967), p. 4-5.

> ...the assumptions on which evaluative designs are based (which are those of experimental design) impose a series of constraints on the evaluator. There can be, for example, no variation in treatment or context once the evaluation is underway, since this would result in the confounding of critical variances. Thus traditional evaluations militate against any concurrent effort at improvement of the treatment and against other simultaneous contextual changes, e.g. the introduction of any other innovation during the term of the evaluation.

Pratt, William F. "Social Research Strategies in Action Programs" Philippine Sociological Review (1966), p. 10

> Where alternative methods or "treatments" are used to introduce a change, evaluation will call for comparative study of the alternative methods. It is in this aspect of action programs that research places the greatest constraint on action... The treatment procedures cannot be altered without risking the validity of evaluation data, for each change becomes itself a part of the treatment and will be practically impossible to evaluate separately.

Short, James F. "Action Research Collaboration and Sociological Evaluation" Pacific Sociological Review (1967), p. 52.

> Large-scale and creative action programs do not "stand still" with practices which appear to be inadequate, or even with those to which they are ideologically committed, but continually probe and shift their strategies to meet problems, old and new.

> Research designs which require a static approach, so as to standardize the "stimulus," are not likely to make much sense to [the practitioners] or to contribute much to cumulative knowledge of social change.

Stufflebeam, Daniel L. "Depth Study of the Evaluation Requirement," Theory into Practice (1966), p. 132.

> While project designs will initially be based on the best knowledge available to the project director, they should be amenable to improvement as the project proceeds. Rigorously controlled evaluation designs should also be avoided, for they usually require that a constant treatment condition be applied in an error-free, laboratory-like context.

Stufflebeam, Daniel L. "The Use and Abuse of Evaluation in Title III," Theory into Practice (1967), p. 128

> ...treatment and control conditions must be held constant throughout the period of the experiment, i.e. they must conform to the initial definitions of these conditions. The Title III or traditional program conditions could not be modified in process, since in that event one could not tell what was being evaluated.
>
> * * * *
>
> ...the application of experimental design to evaluation problems conflicts with the principle that evaluation should facilitate the continual improvement of a program. Experimental design prevents rather than promotes changes in the treatment because treatments cannot be altered in process if the data about differences between treatments are to be unequivocal.

Stufflebeam, Daniel L. and Westerlund, Stuart R. "The Evaluation of Context, Input, Process and Product in Elementary and Secondary Education," BESE, U.S. Office of Education, (Feb. 28, 1967) p. 39.

> To control the environment via experimental design in order to raise the internal validity of information would be self defeating, since change would be stymied rather than accommodated and accelerated by the information collection design.

Wörthen, Blaine R. "Toward a Taxonomy of Evaluation Designs," presented at the AERA 1968 Annual Meeting in Chicago. Evaluation Center, OSU College of Education (mimeo, 1968), p. 4.

> Of course, evaluators as a group are erudite enough to realize that experimental design <u>per se</u> is generally inapplicable in attempts to solve evaluation problems.....

Bibliography

Included here is a brief collection of references to indicate a background of sources and perhaps to guide further inquiry.

Introduction.

Machlup, Fritz, American Economic Review (1965), p. 204.

Part I.

Glassner, L.E., "Transition Room Program," in Evaluation Report 1967 - Volume I. See also 1968 Transition Room Evaluation (both, Pittsburgh Board of Public Education).

Henderson, D.M. and Welty, G.A. "Standardized Testing and Educational Process," working paper prepared for the U.S. Commission on Civil Rights, November 1967.

Westcott, J.H. (ed), An Exposition of Adaptive Control (1962).

Part II.

Office of Research, 1968 Community Utilization Evaluation (Pittsburgh Board of Public Education).

Welty, G.A., "Quality Control, Welfare Economics, and Professor Baier," Journal of Value Inquiry (1967).

Part III.

Grunberg, Emile, "Some Methodological Observations on Macro-Economics," Konjunkturpolitik (1967).

Grunberg, E. and Modigliani, F. "The Predictability of Social Events," Journal of Political Economy (1954).

Simon, H. "Bandwagon and Underdog Effects of Election Predictions," Public Opinion Quarterly (1954). Reprinted as Chapter 5 of Models of Man (1957).

Whitney, H. "The Mathematics of Physical Quantities: Quantity Structures and Dimensional Analysis," American Mathematical Monthly (1968).

Kolmogorov, A.N. and Fomin, S.V. Elements of the Theory of Functions and Functional Analysis Vol. I (1957).

Part IV.

Deming, W.E. Some Theory of Sampling (1950).